Revisiting the Relationship Between Fault Detection, Test Adequacy Criteria, and Test Set Size

Yiqun T. Chen, Rahul Gopinath, Anita Tadakamalla, Michael D. Ernst,
Reid Holmes, Gordon Fraser, Paul Ammann, René Just

@yc_yc_yc_yc

Share your thoughts on this presentation and paper with #ASE2020

W UNIVERSITY of WASHINGTON

CISPA

GEORGE MASON UNIVERSITY

UBC

UNIVERSITÄT PASSAU

How to assess the fault detection capacity of a test set?

Test set adequacy

Test set size

Statement Coverage

Mutation Score

Is test set adequacy a good proxy for fault detection?

Is test set adequacy contributing beyond just size?

Which adequacy measure is the best?

Is test set adequacy correlated with fault detection?*

Usi ? e Defect

And many other papers...!

René Just

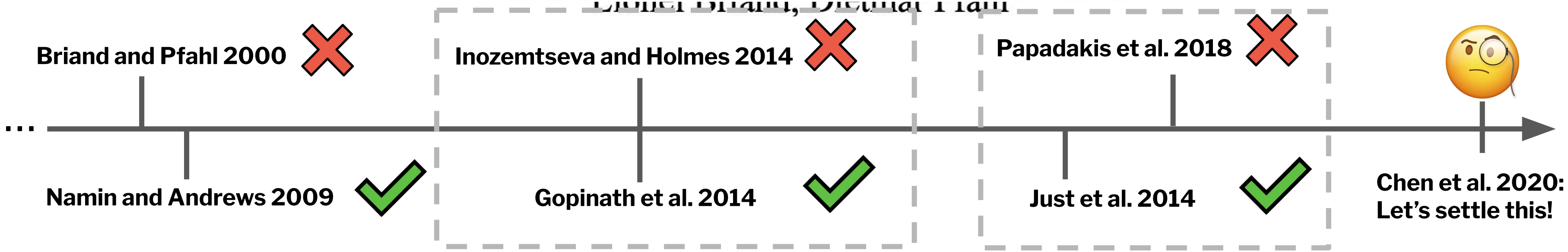
Akbar Siami Namin

Shin Yoo

James H. Andrews

Doo-Hwan Bae

Lionel Briand, Dietmar Pfahl



Briand and Pfahl 2000 ✗

Namin and Andrews 2009 ✓

Inozemtseva and Holmes 2014 ✗

Gopinath et al. 2014 ✓

Papadakis et al. 2018 ✗

Just et al. 2014 ✓

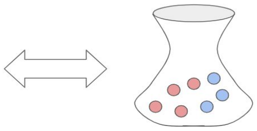
Chen et al. 2020:
Let's settle this!

* Taking test set size into account

Outline

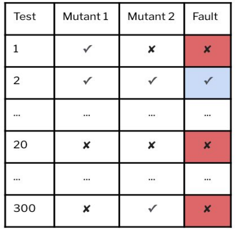
- Review of existing methods

Test	Mutant 1	Mutant 2	Fault
1	✓	✗	✗
2	✓	✓	✓
...	...	...	...
20	✗	✗	✗
...	...	...	...
300	✗	✓	✗

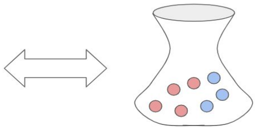


Outline

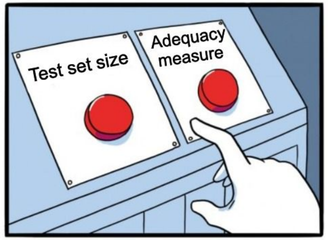
- Review of existing methods



Test	Mutant 1	Mutant 2	Fault
1	✓	✗	✗
2	✓	✓	✓
...	...	...	...
20	✗	✗	✗
...	...	...	...
300	✗	✓	✗

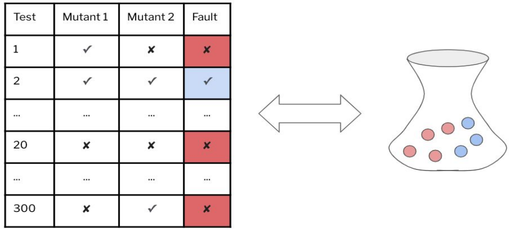


- Ask the right (statistical) question



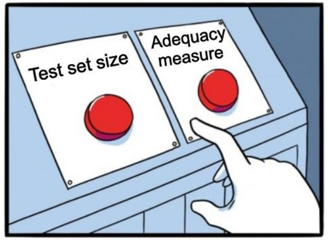
Outline

- Review of existing methods

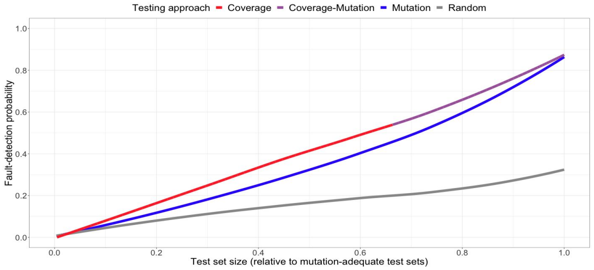


Test	Mutant 1	Mutant 2	Fault
1	✓	✗	✗
2	✓	✓	✓
...	...	...	...
20	✗	✗	✗
...	...	...	...
300	✗	✓	✗

- Ask the right (statistical) question



- Test adequacy measures are valid

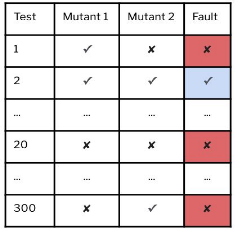


Outline

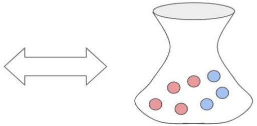
- Review of existing methods

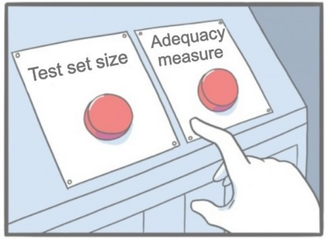
- Ask the right (statistical) question

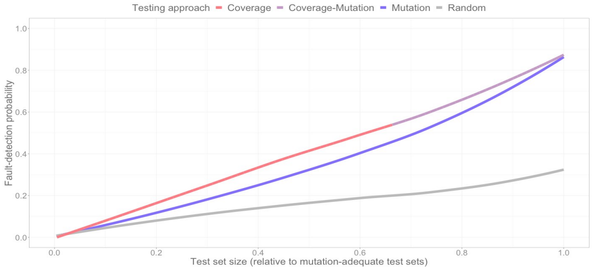
- Test adequacy measures are valid



Test	Mutant 1	Mutant 2	Fault
1	✓	✗	✗
2	✓	✓	✓
...	...	...	...
20	✗	✗	✗
...	...	...	...
300	✗	✓	✗







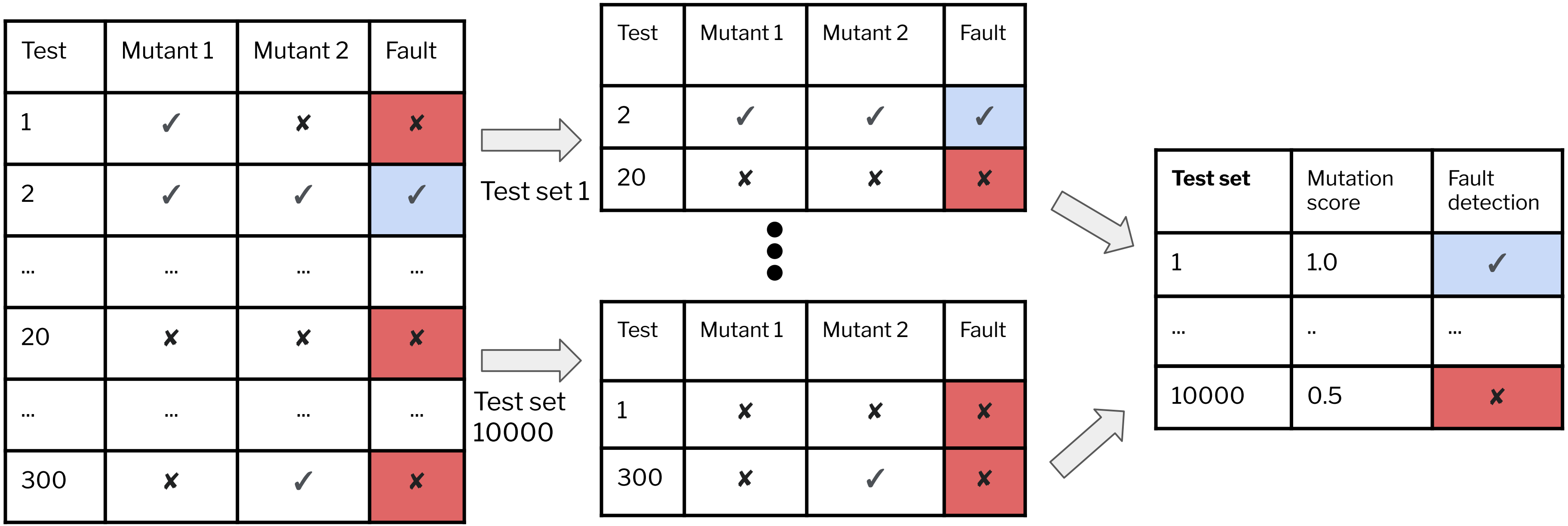
One possible approach: Random selection

Test	Mutant 1	Mutant 2	Fault
1	✓	✘	✘
2	✓	✓	✓
...	...	...	...
20	✘	✘	✘
...	...	...	...
300	✘	✓	✘

- **Random Selection**
  - Generate many test sets by **sampling** from an **existing pool**
  - Focus of our talk

- Alternatives DO exist

Random Selection methodology



Test	Mutant 1	Mutant 2	Fault
1	✓	✗	✗
2	✓	✓	✓
...	...	...	...
20	✗	✗	✗
...	...	...	...
300	✗	✓	✗

Test set 1

Test	Mutant 1	Mutant 2	Fault
2	✓	✓	✓
20	✗	✗	✗

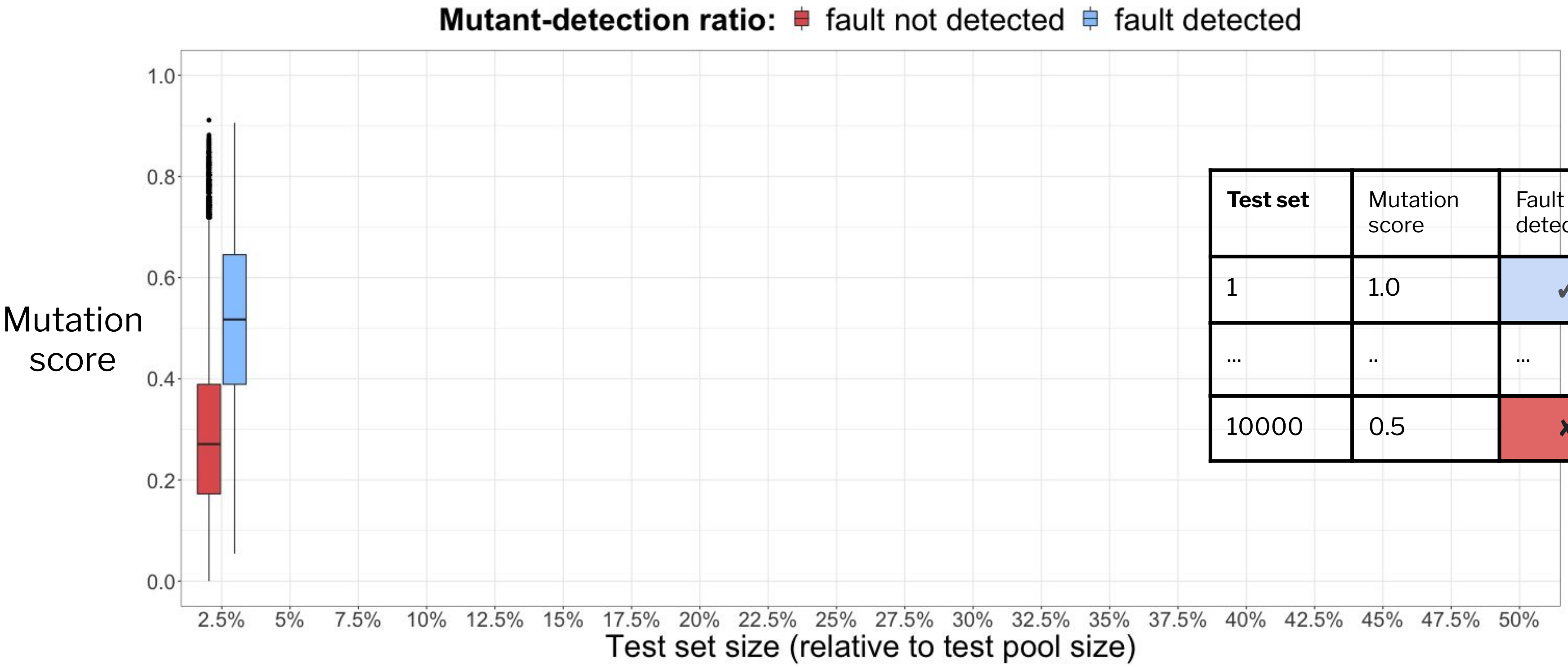
Test set 10000

Test	Mutant 1	Mutant 2	Fault
1	✗	✗	✗
300	✗	✓	✗

Test set	Mutation score	Fault detection
1	1.0	✓
...	..	...
10000	0.5	✗

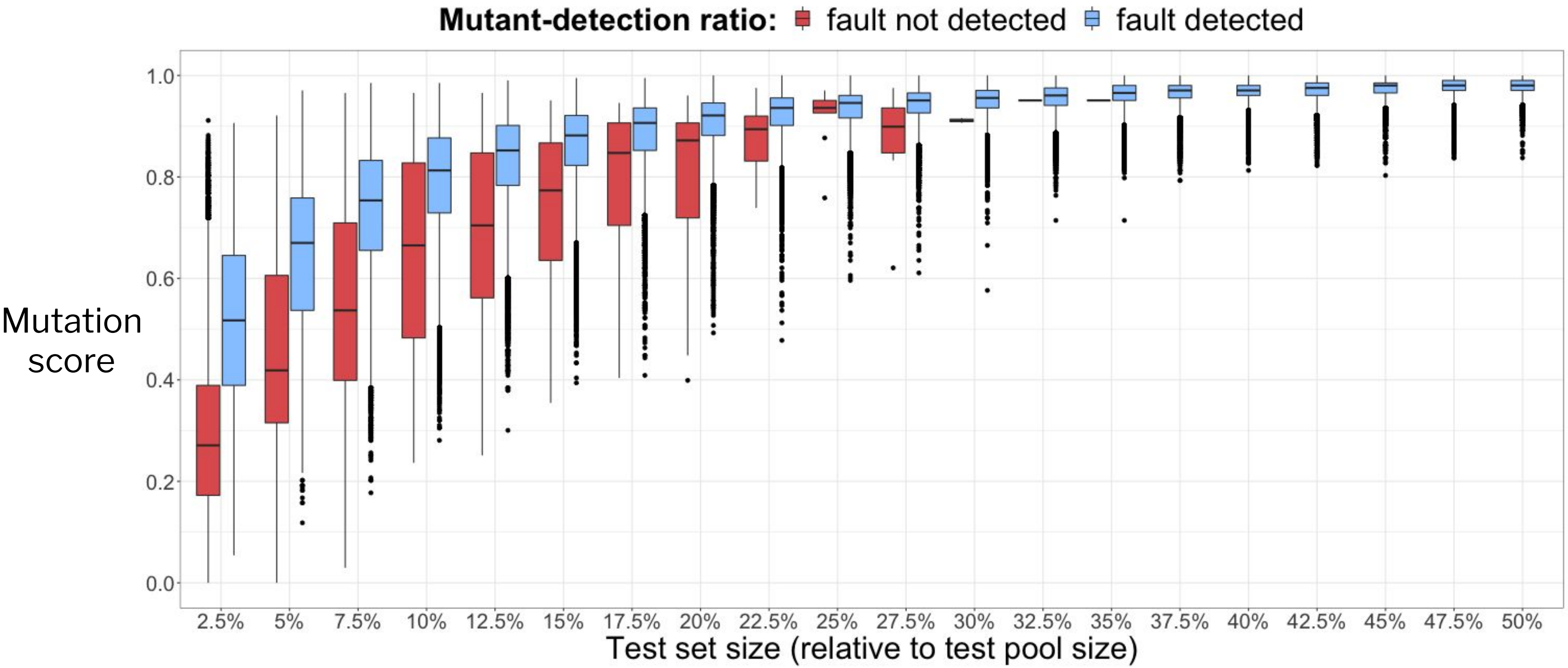
Sample n=2 tests from the test pool **without replacement**, and **analyze** the **results** for **different n**.

Case study: Closure-100 (Defects4J)

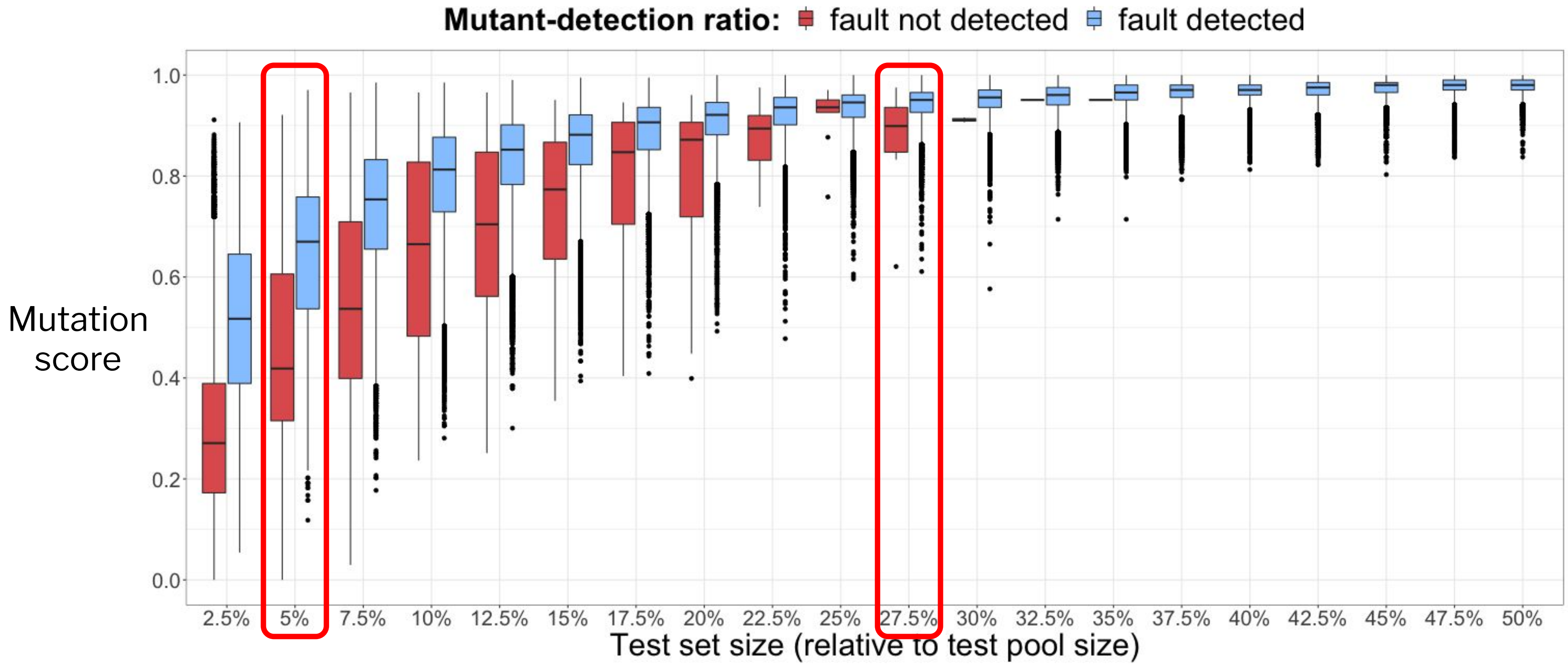


Test set	Mutation score	Fault detection
1	1.0	✓
...	..	...
10000	0.5	✘

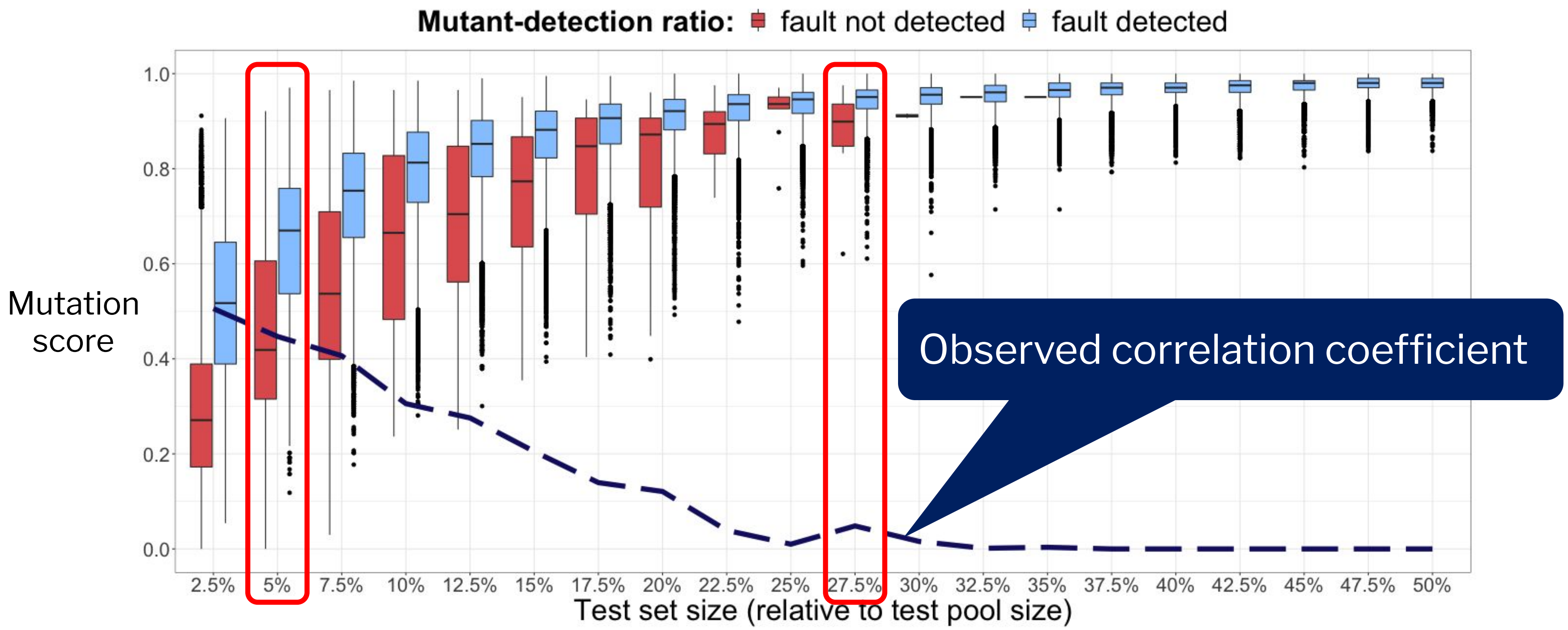
Case study: Closure-100 (Defects4J)



Case study: Closure-100 (Defects4J)



Case study: Closure-100 (Defects4J)



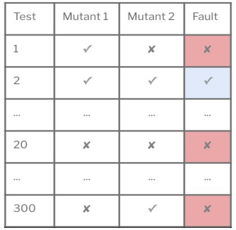
Observed correlation coefficient

Outline

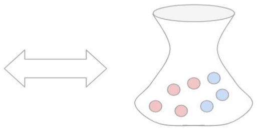
- Review of existing methods

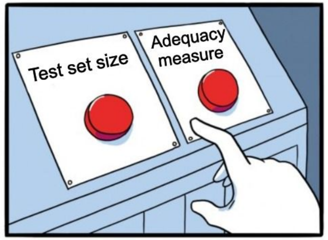
- Ask the right (statistical) question
    - ill-posed question
    - mis-interpretation of correlation

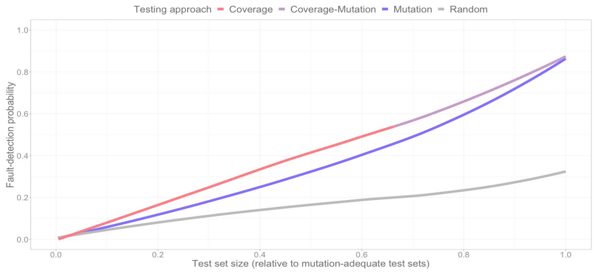
- Test adequacy measures are valid



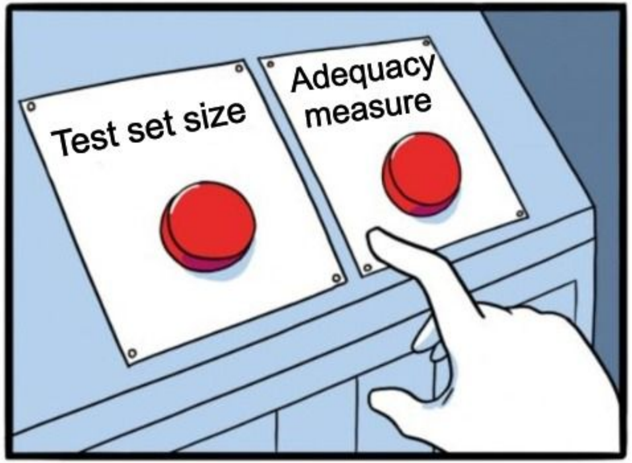
Test	Mutant 1	Mutant 2	Fault
1	✓	✗	✗
2	✓	✓	✓
...	...	...	...
20	✗	✗	✗
...	...	...	...
300	✗	✓	✗







Random selection is prone to misleading conclusions!



An ill-posed question

Q: What are the **individual contributions** of **size and adequacy** to fault detection?

A: Impossible to answer when adequacy and size are **highly correlated**.

- Encode the same information
  - (Hypothetical) adequacy = size

100 x **size** + 0 x adequacy

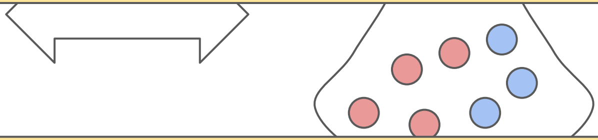
=

0 x size + **100** x **adequacy**

Why does Random Selection fall into this ill-posed question trap?

Test	Mutant 1	Mutant 2	Fault
1			
2			
...	...	...	...
20	✘	✘	✘
...			
300			

Probability of selecting a fault detecting test set
(1) is a **function** of **test set size**, and (2) has an **analytical form**



The same holds for **each mutant!**

Random Selection implies the ill-posed question!

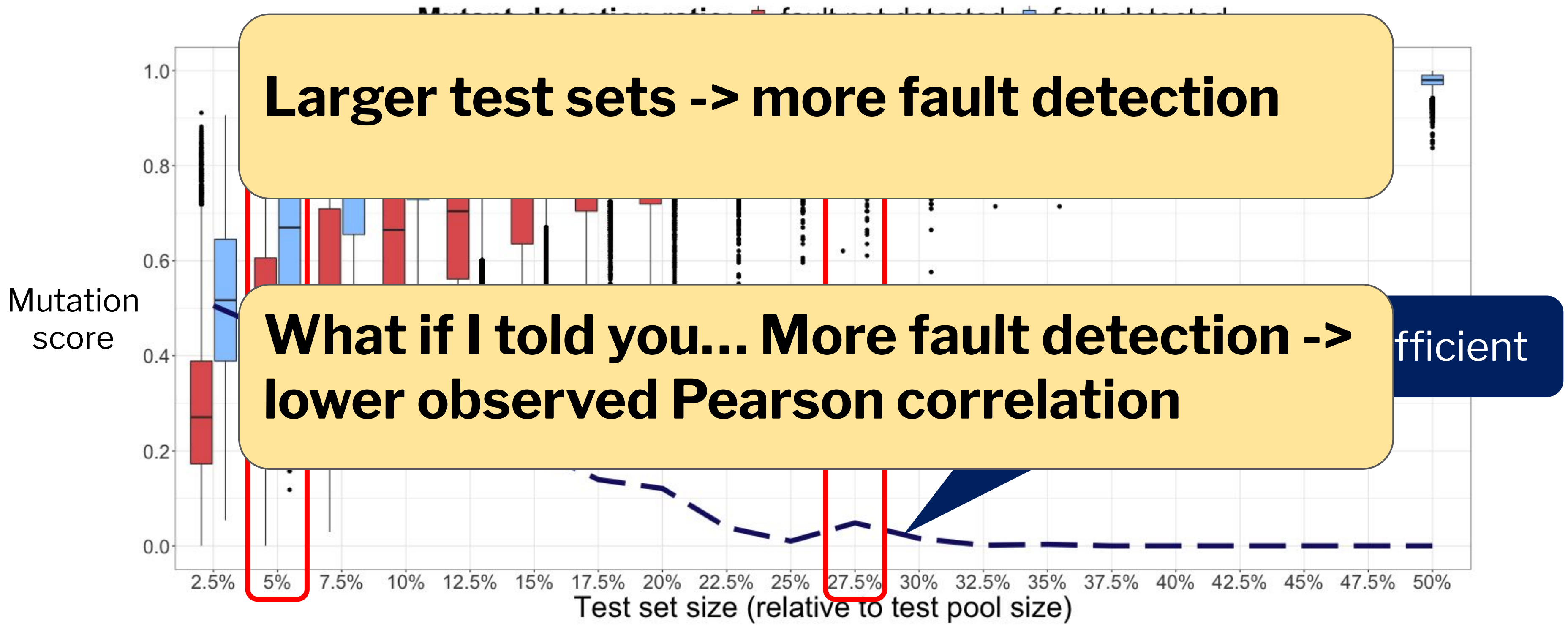
Test			
1			
2	✓	✓	✓
...			
20			
...	...	...	...
300			

Larger test sets -> more fault detection

Larger test sets -> higher mutation score

High pairwise correlation as a result!

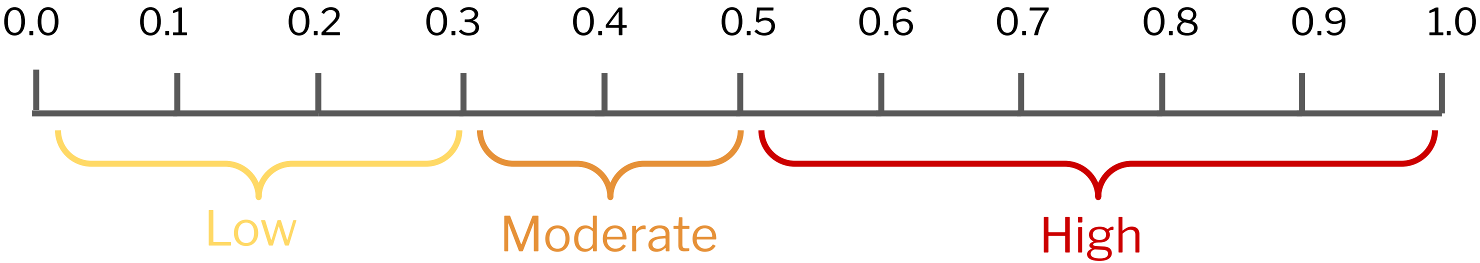
Revisit case study: mis-interpreted Pearson correlation



Larger test sets -> more fault detection

What if I told you… More fault detection -> lower observed Pearson correlation

How we usually interpret Pearson correlation*

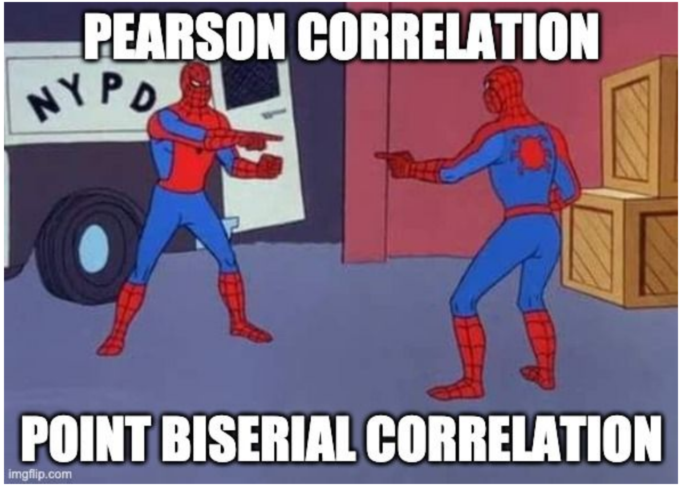


0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1.0

Low

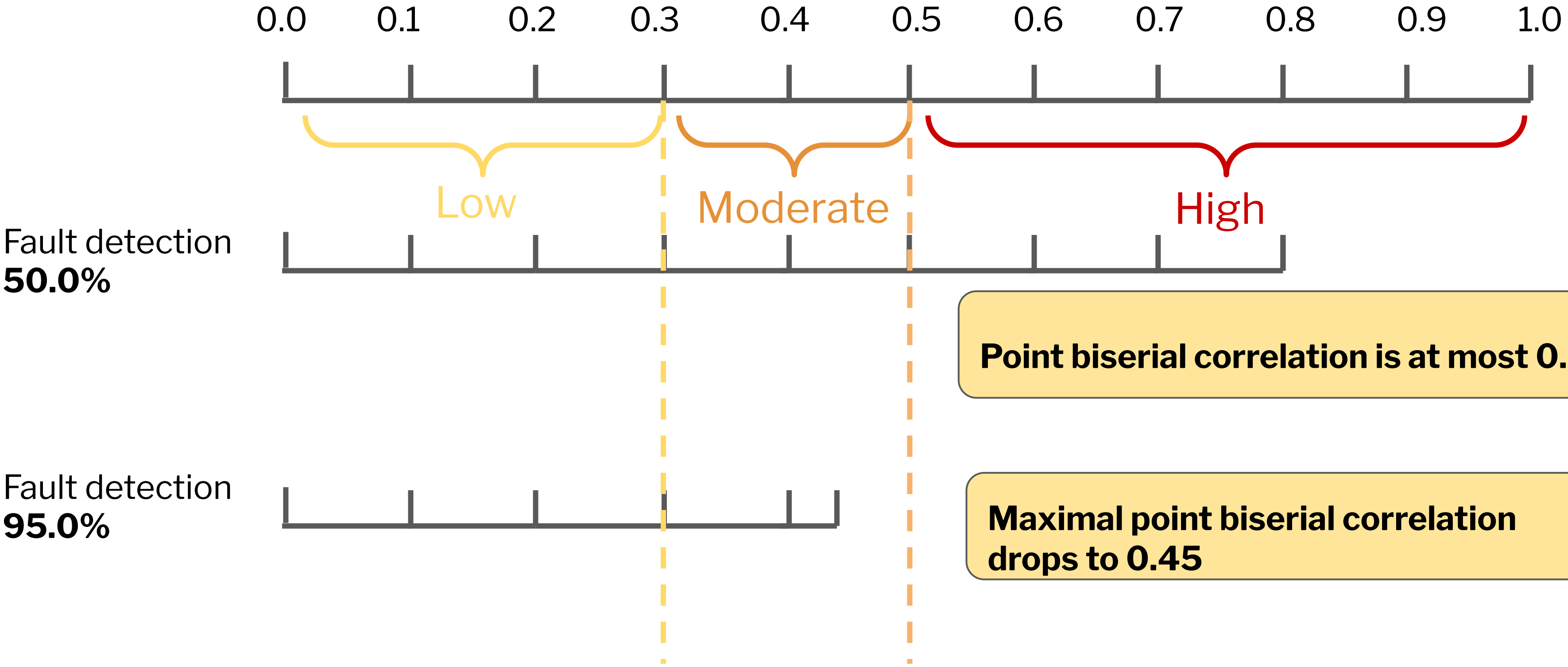
Moderate

High



*Cohen (1988)

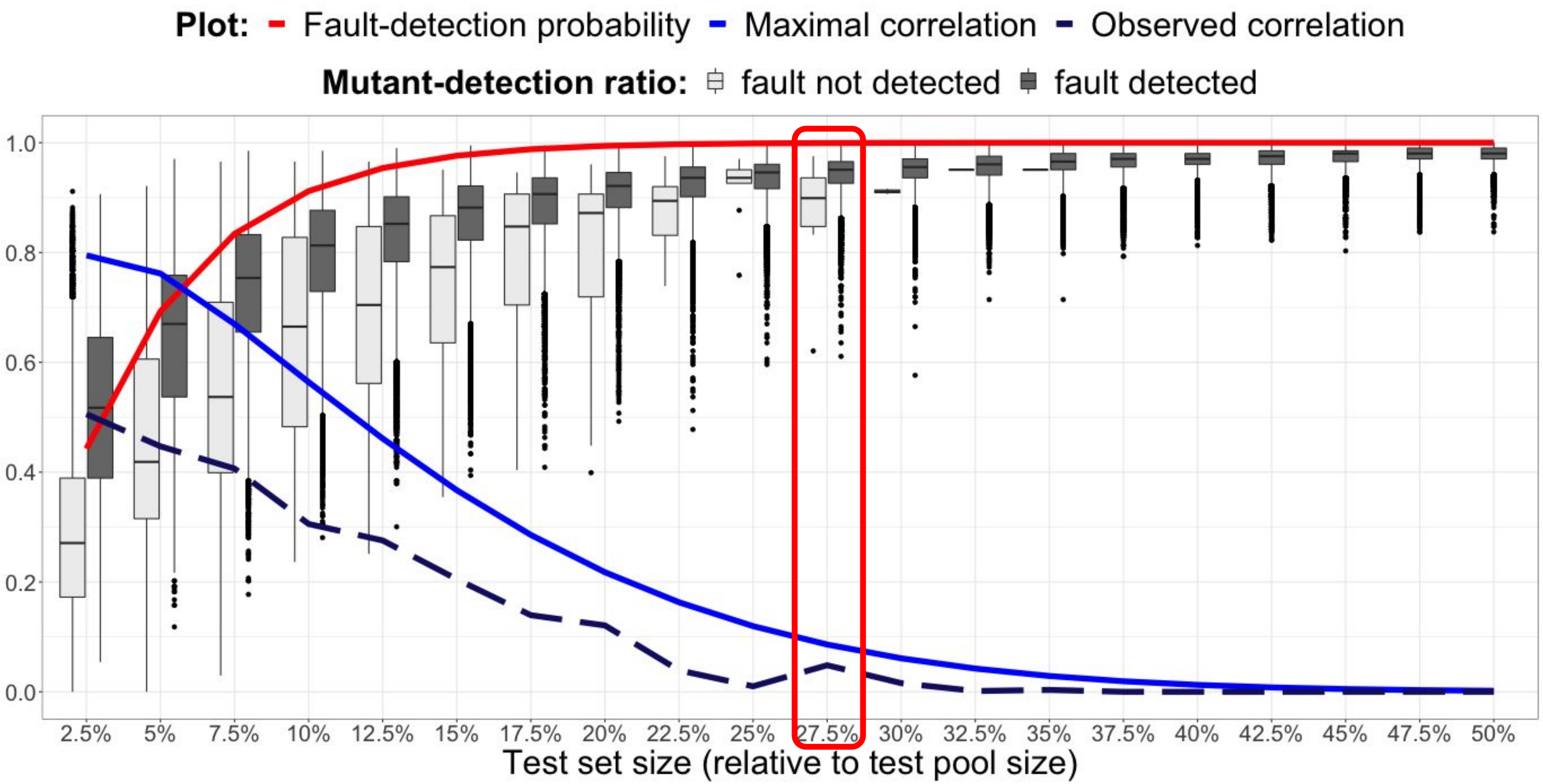
Fun Facts about Point Biserial Correlation



Point biserial correlation is at most 0.8

Maximal point biserial correlation drops to 0.45

Random selection is prone to misleading conclusions!

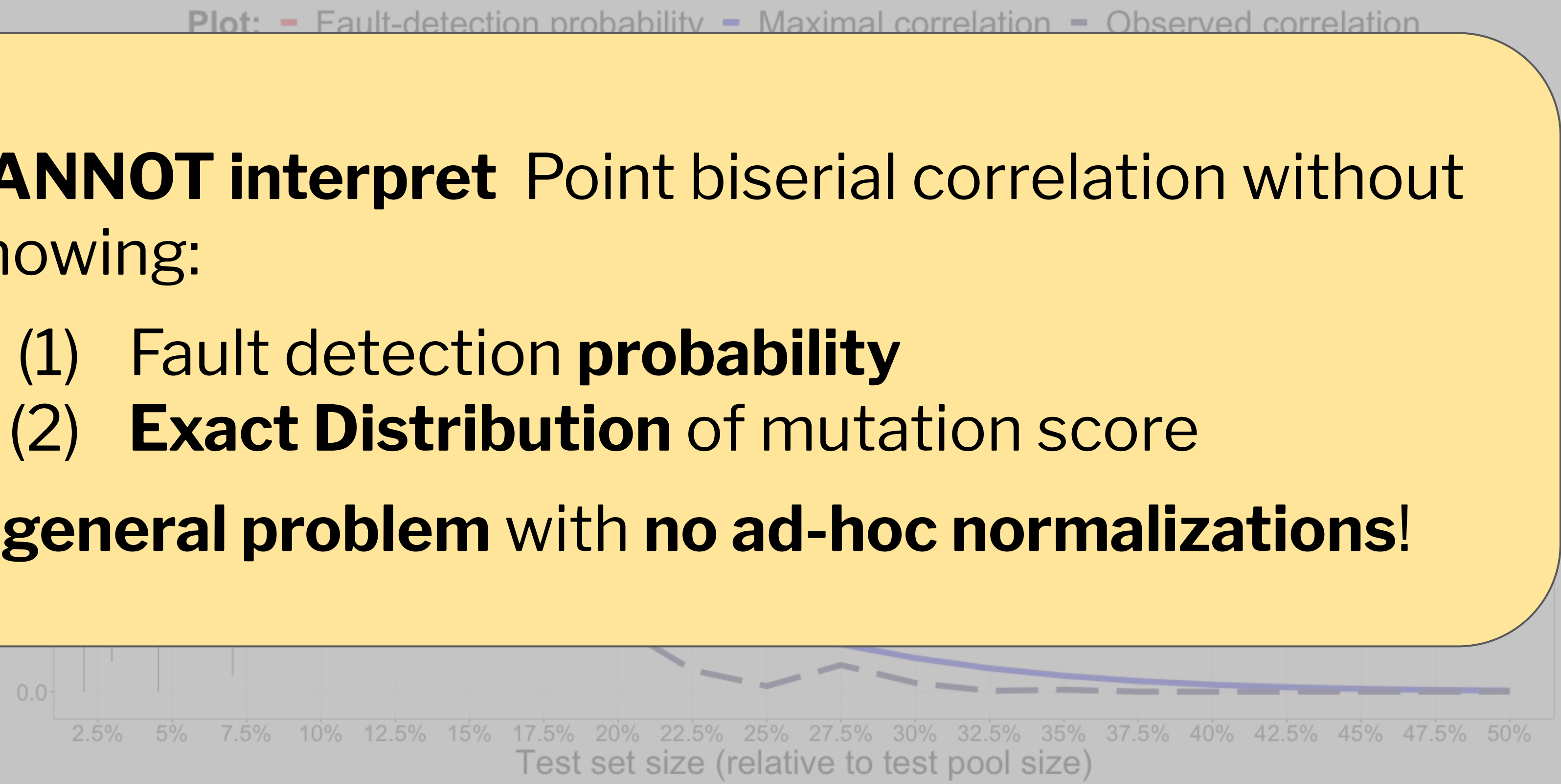


Random selection is prone to misleading conclusions!

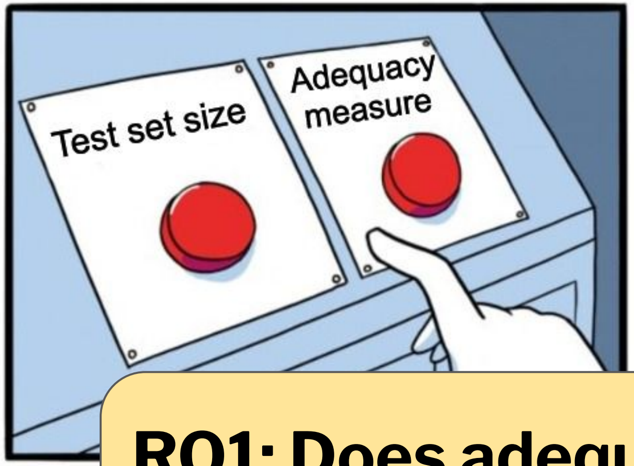
CANNOT interpret Point biserial correlation without knowing:

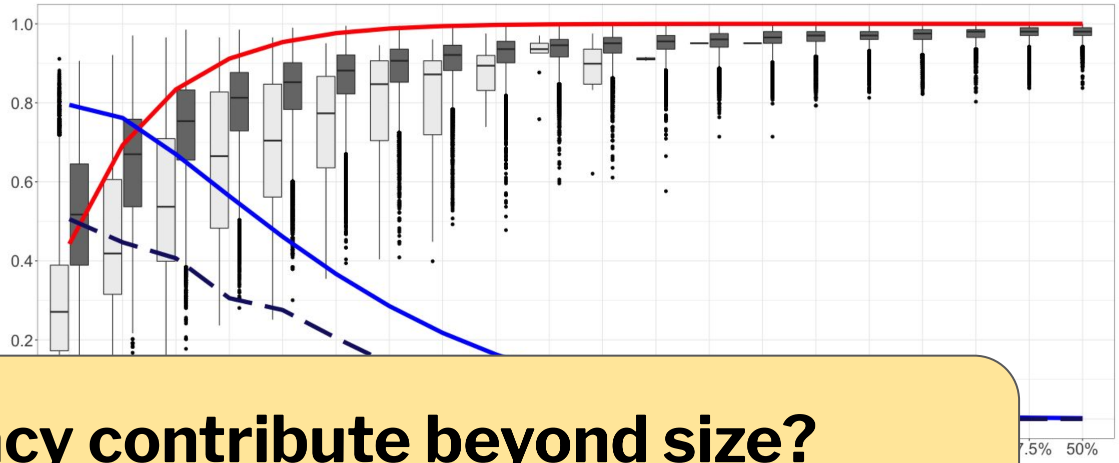
1. (1) Fault detection **probability**
2. (2) **Exact Distribution** of mutation score

A general problem with **no ad-hoc normalizations**!



What can we do to answer our research questions?





RQ1: Does adequacy contribute beyond size?

RQ2: Which adequacy measure is best?

An ill-posed question
correlation doesn't fix that!

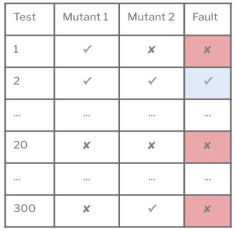
Class imbalance problem
correlation isn't what you think it is!

Outline

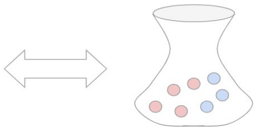
- Review of existing methods

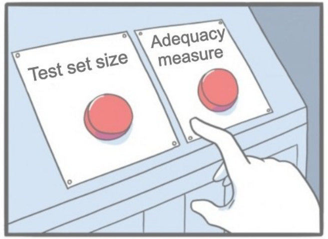
- Ask the right (statistical) question

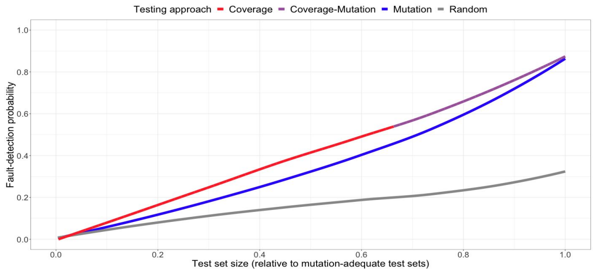
- Test adequacy measures are valid



Test	Mutant 1	Mutant 2	Fault
1	✓	✗	✗
2	✓	✓	✓
...	...	...	...
20	✗	✗	✗
...	...	...	...
300	✗	✓	✗

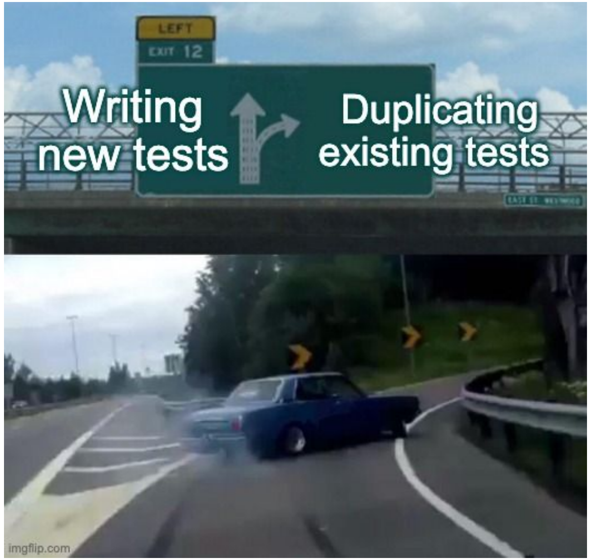






Random Selection is also conceptually flawed!

- Test set size is NOT a meaningful goal in practice!



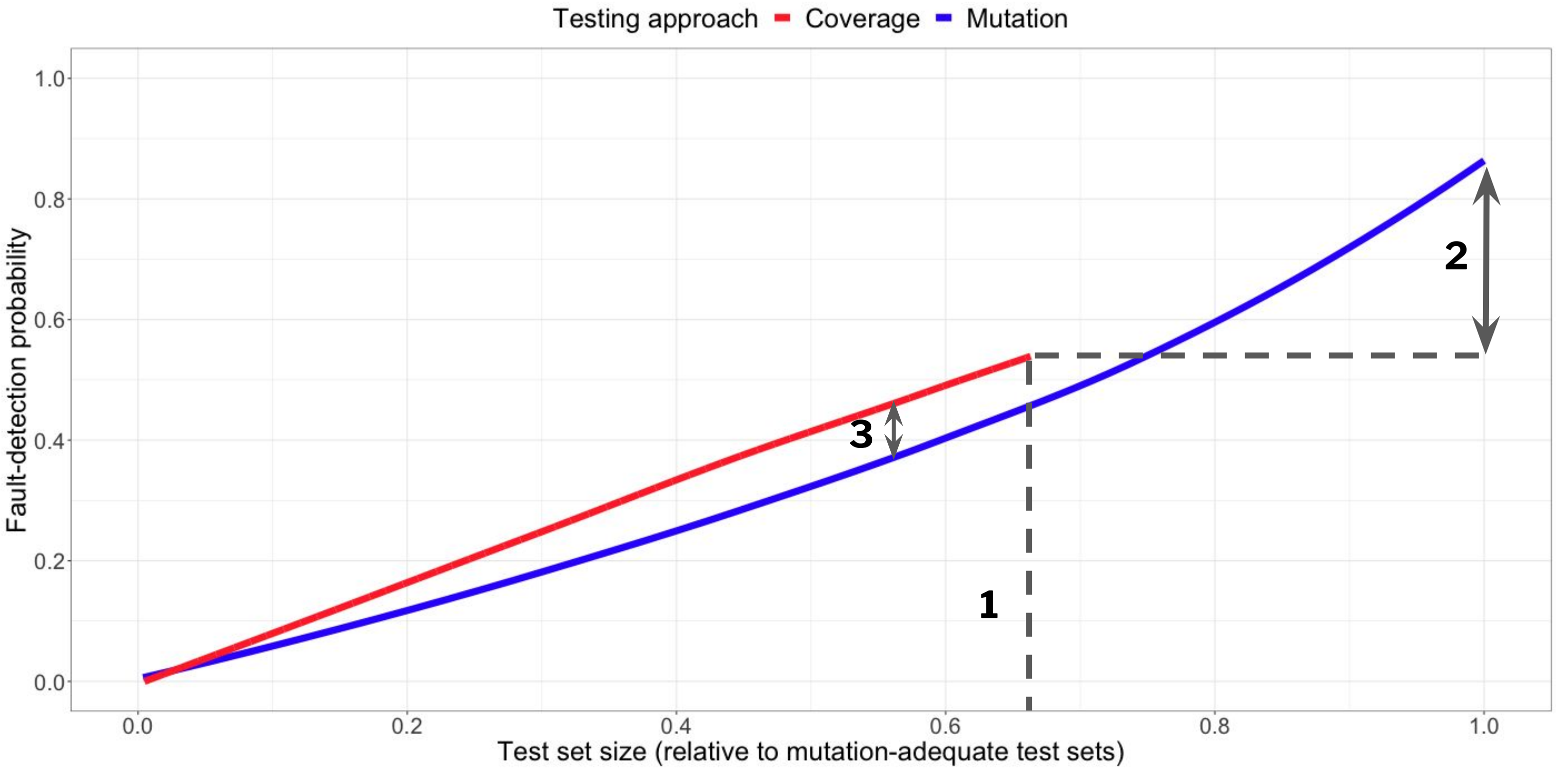
Alternative sets of experiments

- Address the conceptual issue
- Avoid the statistical pitfalls
- Account for test set size

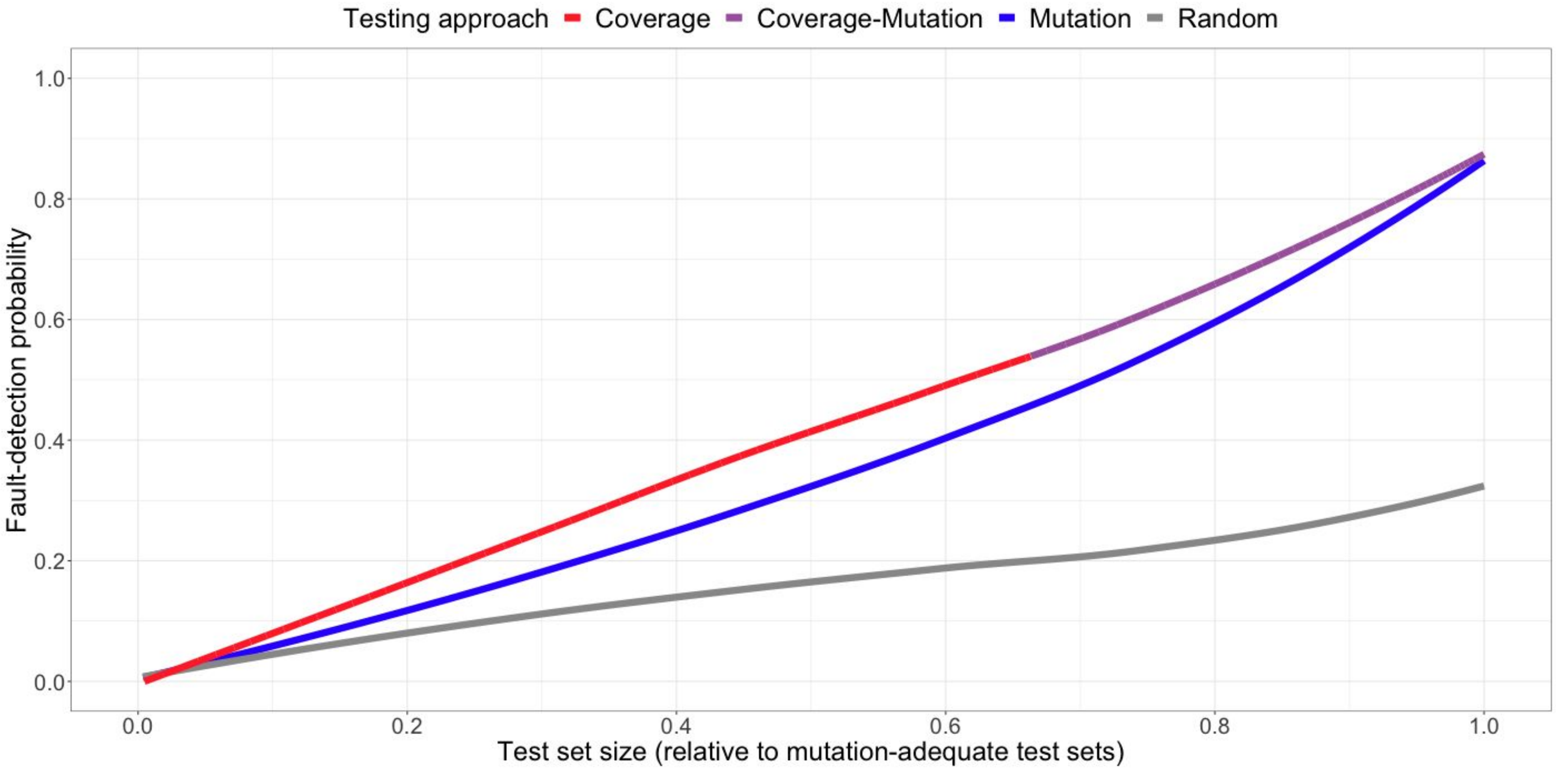
In a nutshell:

- Use adequacy-based testing to achieve a specified level (e.g., 80% coverage)

Statement coverage vs. Mutation score



Statement coverage vs. Mutation score



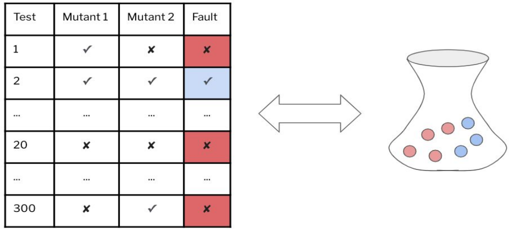
(see also “State of Mutation Testing at Google”, Petrović and Ivanković (2018))

Conclusions

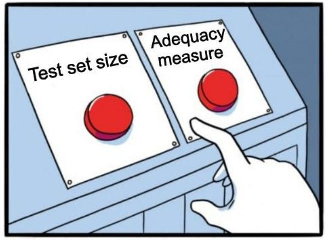
- Random selection is prone to misleading results.

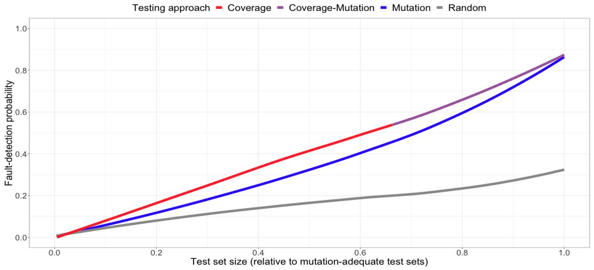
- Mutation & coverage are VALID adequacy measures and contribute beyond just size.

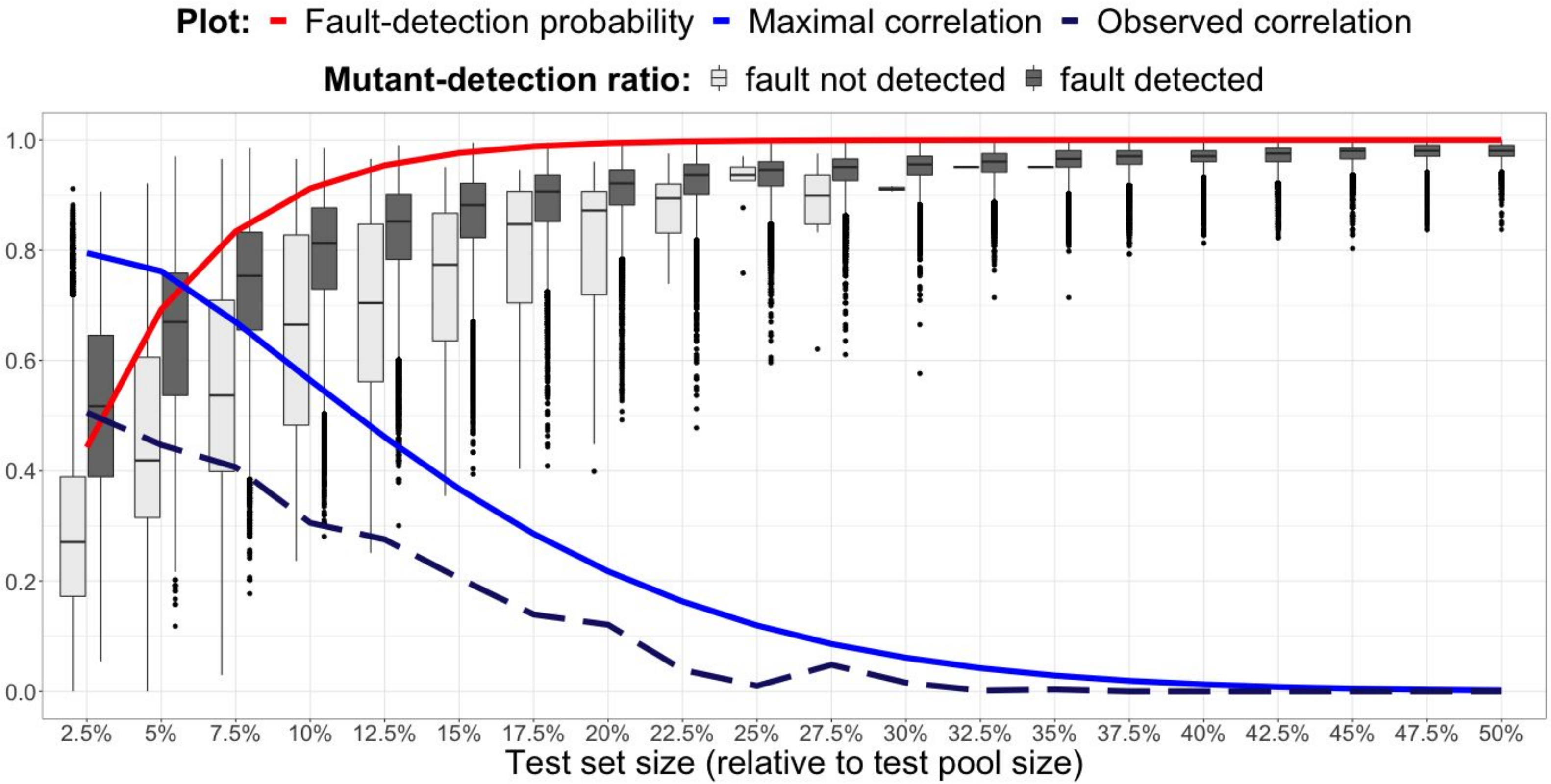
- Want effective tests? Coverage + Mutation



Test	Mutant 1	Mutant 2	Fault
1	✓	✗	✗
2	✓	✓	✓
...	...	...	...
20	✗	✗	✗
...	...	...	...
300	✗	✓	✗






Plot: ▬ Fault-detection probability ▬ Maximal correlation ▬ Observed correlation
Mutant-detection ratio: ⊟ fault not detected ■ fault detected
1.0
0.8
0.6
0.4
0.2
0.0
2.5%
5%
7.5%
10%
12.5%
15%
17.5%
20%
22.5%
25%
27.5%
30%
32.5%
35%
37.5%
40%
42.5%
45%
47.5%
50%
Test set size (relative to test pool size)